EVIDENCE-BASED CASE REVIEW

Which guidelines can we trust?

Assessing strength of evidence behind recommendations for clinical practice

Alessandro Liberati
Istituto di Statistica Medica, Università degli Studi di Modena e Reggio Emilia, Modena, and Milan, Italy
Roberto Buzzetti
Modena
Roberto Grilli
Milan and Rome
Nicola Magrini
Modena
Silvia Minozzi
Milan

Correspondence to:
Dr Liberati
a.liberati@ausl.mo.it

Competing interests:
None declared

West J Med
2001;174:262-265

The parents of a healthy, asymptomatic 5-year-old boy are anxious about his health and ask about the appropriateness of undergoing a screening examination with urinalysis. You search for existing recommendations on this topic and find the book, *Putting Prevention Into Practice*.[1] You find the 2 statements outlined below.

- American Academy of Family Physicians and US Preventive Services Task Force:
  Routine screening of males and most females for asymptomatic bacteriuria is not recommended. The Canadian Task Force on the Periodic Health Examination and the US Preventive Services Task Force recommend against screening for asymptomatic bacteriuria with urinalysis in infants, children, and adolescents.
- American Academy of Pediatrics:
  Urinalysis should be performed once at 5 years of age. Also, dipstick leukocyte esterase testing to screen for sexually transmitted diseases should be performed once in adolescence, preferably at 14 years of age.

This clinical scenario raises a number of important questions:

- What sort of evidence has been used to come to these different conclusions?
- How have the 2 committees looked at and appraised the evidence?
- Have they used an explicit approach to classify the quality of existing studies?
- If they have, indeed, used an explicit approach, which elements—such as study design, study conduct, or relevance of the outcome measures—have they considered?

Explicit recommendations for clinical practice, such as guidelines or diagnostic and therapeutic protocols, are published frequently, but many have conflicting recommendations. To decide which guidelines we should follow, we need common criteria to assess the quality of available evidence. Although it is generally agreed that practice guidelines should explicitly assess the quality of the evidence that supports different statements, this is still uncommon.[2]

Historically, the Canadian Task Force was the first to attempt to classify levels of evidence supporting clinical recommendations. It did this by reviewing the indications for preventive interventions and producing recommendations with an explicit grading of the supporting evidence.[3] These were subsequently adopted by the US Preventive Services Task Force.[4] The original approach used by the Canadian Task Force classified randomized controlled trials (RCTs) as the highest level of evidence, followed by non-RCTs, cohort and case-control studies (representing fair evidence), comparisons among times and places with or without the intervention, and at the lowest level, "expert opinion." This approach is simple to understand and easy to apply, but it implicitly assumes that RCTs, no matter how small or large or how properly conducted, always produce better evidence than nonexperimental studies such as cohort or case-control studies. This approach also ignores the issue of heterogeneity and, thus, what to do when results from several RCTs or other nonexperimental studies vary.

Other scales proposed since that of the Canadian Task Force still rely on methodologic design of primary studies as the main criterion. These have incorporated systematic reviews and meta-analyses, which are placed above RCTs in the "hierarchy of evidence." Whereas this allows for a possibly more refined grading of levels of evidence, it suffers from the same limitation—ie, that attention is given to the a priori validity of the methods used. More recently, scales assessing the quality of study conduct and

Summary points

- An assessment of the strength of evidence that underlies recommendations or guidelines may help clinicians decide which to follow
- Evidence may be graded on the basis of a priori validity of study design, aspects of study conduct, consistency of evidence, or clinical relevance
- Many scales are available to assess strength of evidence, although none is wholly satisfactory
- The strength of recommendations should depend on the strength of evidence, judgment about values to be ascribed to various outcomes, and contextual issues such as availability of resources and effects on other services

See this article on our web site for links to other articles in the series.

the consistency of results across different studies have been proposed.

The aims of this article are as follows:

- to review existing scales aimed at assessing the quality of evidence supporting treatment recommendations
- to discuss the need to go beyond the assessment of methodologic quality—whether measured a priori by looking at study design or a posteriori by looking at study conduct—to include an explicit assessment of the epidemiologic and clinical relevance of the evidence
- to suggest which direction research in this area should take.

We will not address how strength of recommendations has been assessed. This is a complex concept that implies value judgments and an explicit methodologic assessment of available studies. As recently suggested (A Oxman, S Flottorp, J Cooper, et al, "Levels of Evidence and Strength of Recommendations," unpublished data, 1999), "strength of recommendations" is a construct that should go beyond levels of evidence to incorporate more subjective considerations, such as patient- or setting-specific applicability; tradeoffs among risk, benefits, and costs; and the like.

WAYS TO CLASSIFY LEVELS OF EVIDENCE

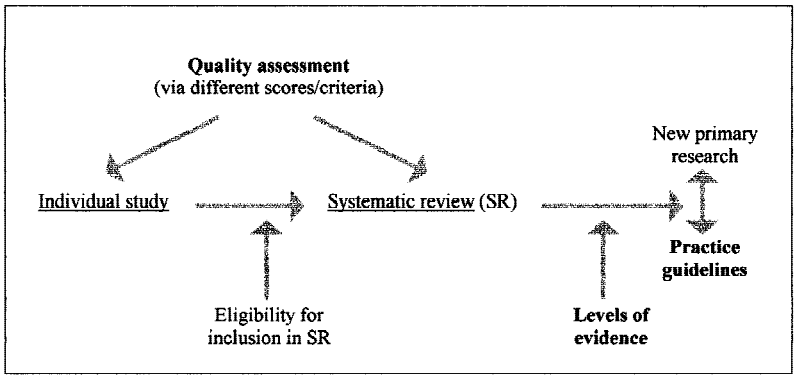
When used for individual studies, quality assessment provides explicit criteria to separate valid from invalid studies (usually referred to as "internal or scientific validity"). When used in a systematic review, quality assessment can assist in qualifying the recommendations to be incorporated into practice guidelines or recommendations (figure).

A priori validity of study design

The validity of study design is the oldest and still most commonly used approach to levels of evidence classification. The 2 main advantages of this approach are its explicit nature and the fact that a general consensus exists regarding the hierarchy of different types of study designs in their ability to prevent bias.[3,4] On the other hand, this approach relies exclusively on issues of design, thereby ignoring issues of study conduct and of the consistency and clinical and epidemiologic relevance of study findings.

Quality of study conduct

Despite its appeal, the feasibility of analyzing the quality of the conduct of the study is seriously jeopardized by the lack of consensus regarding the appropriate indicators of study validity (lack of an agreed-on gold standard). Not even for RCTs—the most standardized type of study design—is there an agreement on whether a quality score or a criteria-based system is better.[5] Several years ago, Emerson et al[6] failed to demonstrate the predictive validity of a widely used, detailed method for quantifying the quality of trials, which included evaluating adequacy of descriptions, blinding, and essential measurements. More recently, Juni et al[7] reported substantial differences in the assessed "quality" of an article, depending on the method used to measure it. Thus far, the only item for which there is clear empiric evidence of bias prevention is the quality of



Where quality assessment can be carried out and classification of levels of evidence made explicit

Matazzi/SPL

Should a healthy 5-year-old child be screened with urinalysis?

the randomization process, defined as the extent to which the allocation process was concealed.[8]

Consistency of results across studies

Consistency of results is an important issue, although it must be adjusted for the study design and quality of study conduct. Dramatically large effects may be consistently reported in studies of lower methodologic quality (eg, a series of observational studies), but further tests based on more rigorous designs may then indicate much smaller, if any, effect.[9] Relative to the quality of study conduct, consistency per se does not imply validity, as a series of individual studies can be systematically wrong if the same biases exist (such as in selecting the study population or using systematically inaccurate measurements).

Clinical relevance of study results

The difficulty with ensuring clinical relevance of results is in defining generic criteria for relevant end points of interventions across diseases or conditions and the likely dependence of the judgment(s) from the perspective of the assessor(s)—ie, patient, provider, or purchaser.

EXISTING SCALES FOR CLASSIFYING LEVELS OF EVIDENCE

The table lists 9 scales available to assess levels of evidence.[3,4,10-16] All scales explore the dimension of a priori study validity, but the level of details varies from the simplest approach of the Canadian Task Force (4 levels) to the more complex and analytic taxonomy proposed by more recent scales. Only 4 scales also critically appraised the quality of the study conduct, through predefined criteria, although they differed on criteria applied and operational definitions.[13-16] Consistency of results is incorporated into 4 scales.[12,14-16] However, heterogeneity is neither clearly nor consistently defined across scales.

Some scales, such as the Canadian Task Force and the US Preventive Services Task Force, separate levels of evidence from strength of recommendations. In the case illustrated in the opening paragraph, for example, the evidence for the use of routine urinalysis was level I, and the recommendation was "type E" (do not perform), but in others, the 2 are more closely tied.

The state of the art is still, therefore, unsatisfactory. Although 3 scales look at all 3 dimensions listed in the table,[14-16] the main challenge for a better approach to levels of classifying evidence is how to combine the 3 dimensions outlined earlier with the clinical and epidemiologic relevance of the study findings.

*Dimensions of quality explored by different scales**

Scale and study	No. of levels	Study design	Quality of conduct	Consistency of results
Canadian Task Force, 1990[3]	4	X		
US Preventive Services Task Force, 1996[4]	5	X		
AHCPR, 1992[11]	5	X		
Guyatt et al., 1995[12]	6	X		X
Eccles et al, 1996[10]	6	X		
Hadorn et al, 1996[13]	7	X	X	
Ball et al, 1997[14]	10	X	X	X
Liddle et al, 1997[15]	5	X	X	X
Jovell et al, 1997[16]	9	X	X	X

AHCPR = Agency for Health Care Policy and Research.
*An "X" indicates an area explored; a blank space indicates it was not addressed by the study.

NEED TO CONSIDER EPIDEMIOLOGIC AND CLINICAL RELEVANCE

When the Canadian Task Force scale was originally proposed, RCTs were less common and requirements for drug approval were less stringent, so that evidence from such trials was often not available. With the much wider availability of these trials, the scales have become insensitive to differences in the quality of supporting evidence. As a result, it may be inappropriate to accept the presence of 1 or 2 RCTs as sufficient evidence in favor of an intervention.

Critically appraising aspects of the question addressed is also important: was the study designed to explore long-term versus short-term use of the treatment, the type of skill or experience required by the providers, and the availability of the appropriate level of care? Two issues are central here: the nature of the end point (whether it is hard or soft, clinical versus surrogate, and what its relationship is to the quality or quantity of life), and the appropriateness of the comparator chosen (whether different candidate interventions are directly compared, or are they each only compared with nothing or placebo).

Strong evidence of effect for an intervention does not necessarily translate into equally strong recommendations for its use. Cost, the values placed on the outcomes by physicians and patients, and feasibility must all be factored into recommendation, along with the evidence (strong or otherwise). For instance, when assessing the evidence for and against breast cancer screening on a population level, although the evidence of effectiveness is strong (the usefulness of mammography screening in women >50 years is supported by several RCTs), it may still be inappropriate to recommend screening if the other criteria for implementation are not met. For example, too few well-trained radiologists may be available to read the mammograms, pathologists to interpret the biopsy specimens, or surgeons

to perform appropriate surgery in a particular health district. On the other hand, evidence that is less strong may lead to strong recommendations when there are no viable alternatives and the do-nothing approach is not feasible.

CONCLUSIONS AND FUTURE DIRECTIONS

Although more recent scales take into account the quality of study conduct, we found no scale that explicitly includes the clinical and epidemiologic relevance of the question addressed by the studies. The use of only methodologically based quality assessment to judge the evidence supporting an intervention is inadequate, especially in an area of therapy where RCTs (ie, the highest methodologic level of evidence) are commonly available.

A possible solution is to abandon the idea that a generic scale can satisfactorily assess levels of evidence for a particular therapeutic or diagnostic question. A generic scale could be integrated with specific criteria targeted to the nature of the question being explored. The generic scale should look at the a priori quality of study design (ie, has the appropriate design for the question at issue been used?) and at the validity of the study conduct. Scales such as those discussed by Hadorn, Ball, Liddle, and Jovell and their co-workers[13-16] are all good steps in this direction, although an effort to provide operational definitions is needed. The criterion-specific items might concentrate on the relevance of the end point and on the appropriateness of its timing, setting, and level of care.

What is the lesson for the decision to be taken in the clinical scenario at the start of this article? Going back to the original sources, you find that the US Preventive Services Task Force report indicates that evidence from both RCTs and observational studies support the recommendation not to perform a screening test for asymptomatic bacteriuria in infants, children, and adolescents.[4] The recommendation by the American Academy of Pediatrics is simply an unqualified consensus statement without any reference to the level of evidence supporting it.[17] Despite the limitations of existing scales available to assess levels of evidence, having an explicit approach for ranking the methodologic quality of available studies is useful, at least for the time-being. It is particularly helpful when comparing different recommendations allegedly drawn from the same type of evidence.

Authors Alessandro Liberati is affiliated with the Istituto di Statistica Medica, Università degli Studi di Modena e Reggio Emilia. Alessandro Liberati, Roberto Buzzetti, and Nicola Magrini are affiliated with the Centro Valutazione Efficacia Assistenza Sanitaria, Modena; and Alessandro Liberati and Roberto Grilli are associated with the Centro Cochrane Italiano Istituto "Mario Negri," Milan, Italy. Dr Grilli is also with the Agenzia Servizi Sanitari Regionali in Rome.

This article was edited by Virginia A Moyer of the department of pediatrics, University of Texas Medical Center at Houston. Articles in this series are based on chapters from Moyer VA, Elliott EJ, Davis RL, et al, eds. *Evidence-Based Pediatrics and Child Health.* London: BMJ Books; 2000.

References

1 *Putting Prevention Into Practice: Clinician's Handbook of Preventive Services: Children and Adolescents—Screening.* 2nd ed. US Dept of Health and Human Services; 1998.
2 Grilli R, Magrini N, Penna A, Mura G, Liberati A. Practice guidelines developed by specialty societies: the need for a critical appraisal. *Lancet* 2000;355:103-106.
3 Woolf SH, Battista R, Anderson GM, et al. Assessing the clinical effectiveness of preventive maneuvers: analytic principles and systematic methods in reviewing evidence and developing clinical practice recommendations. A report by the Canadian Task Force on the Periodic Health Examination. *J Clin Epidemiol* 1990;43:891-905.
4 US Preventive Services Task Force. *Guide to Clinical Preventive Services.* 2nd ed. Baltimore: Williams & Wilkins; 1996.
5 Moher D, Jadad AR, Tugwell P. Assessing the quality of randomized controlled trials: current issues and future directions. *Int J Technol Assess Health Care* 1996;12:195-208.
6 Emerson JD, Burdick E, Hoaglin DC, Mosteller F, Chalmers TC. An empirical study of the possible relation of treatment differences to quality scores in controlled randomized trials. *Control Clin Trials* 1990;11:339-352.
7 Juni P, Witschi A, Bloch R, Egger M. The hazards of scoring the quality of clinical trials for meta-analysis. *JAMA* 1999;282:1054-1060.
8 Schulz KF, Chalmers I, Hayes RJ, Altman DG. Empirical evidence of bias: dimensions of methodological quality associated with estimates of treatment effects in controlled trials. *JAMA* 1995;273:408-412.
9 Sacks H, Chalmers TC, Smith H Jr. Randomized versus historical controls for clinical trials. *Am J Med* 1982;72:233-240.
10 Eccles M, Clapp Z, Grimshaw J, et al. North of England evidence based guidelines development project: methods of guideline development. *BMJ* 1996;312:760-762.
11 *Acute Pain Management.* Rockville, MD: US Dept of Health and Human Services, Public Health Services, Agency for Health Care Policy and Research; 1992. AHCPR publication 92-0038.
12 Guyatt GH, Sackett DL, Sinclair JC, Hayward R, Cook DJ, Cook RJ. Users' guides to the medical literature: IX, a method for grading health care recommendations: Evidence-Based Medicine Working Group. *JAMA* 1995;274:1880-1884 [published erratum appears in *JAMA* 1996;275:1232].
13 Hadorn DC, Baker D, Hodges JS, Hicks N. Rating the quality of evidence for clinical practice guidelines. *J Clin Epidemiol* 1996;49:749-754.
14 Ball C, Sackett D, Phillip B, Straus S, Haynes B. *Levels of Evidence and Grades of Recommendations.* Oxford, UK: Centre for Evidence based Medicine; 1997.
15 Liddle J, Williamson M, Irwig L. *Method for Evaluating Research and Guideline Evidence.* Sidney, Australia: NSW Health Department; 1997.
16 Jovell AL, Navarro-Rubio MD. Evaluacion de la evidencia cientifica. *Med Clin (Barc)* 1997;105:740-743.
17 American Academy of Pediatrics. Recommendations for preventive pediatric health care. *Pediatrics* 1995;96:373-374.